fMRI analysis using Support Vector Machines

Bachelor-Thesis

submitted on: November 26, 2012

to the School of Busines and Economics
of Humboldt Universität zu Berlin
in particular fulfillment of the requirments for the Degree of
Bachelor of Science

Name:	Patrik Bey
Matrikelnr.:	524469
Chair:	Ladislaus von Bortkiewicz Chair of Statistics
Field of study:	Betriebswirtschaftslehre
Supervisor:	Prof. Dr. W. Härdle

Contents

List of Figures

List of Tables

Abstract
Support Vector Machines (SVM) as a tool has become one of the most established techniques for analyzing functional magnetic resonance imaging (fMRI) data in recent years. The ability to deal with very high dimensions in the feature space as well as it's robustness have played a crucial role in promoting SVM's popularity among scientists in the field of neuroscience and related research. These data were acquired during an experiment conducted by the Max Planck Institue where 22 subjects were given an investment decision task with changing levels of uncertainty. Recent literature suggests that a lot of information about individual differences in decision making lies in the variability of the blood-oxygen-level dependent (BOLD) fMRI signals. Given the computed variability of the BOLD level following the stimuli I train an SVM to classify the subjects with respect to their risk attitude. By reducing the dimensions of the input to the areas of the brain previously ascertained as relevant for decision making under uncertainty I decrease the computation time without using time intensive dimension reduction techniques. I then compare my results with the results and technique presented by **Mohr and Härdle et al. (2010)**.

Keywords: SVM, fMRI, classification, decision making, decision under risk

1 Introduction

These methodological developments are providing cognitive neuroscientists with the opportunity of tackling new research questions that are relevant for our understanding of the functional organization of the brain [7]. In neuroscience, decision making gains more and more interest as a field of research, where interdisciplinary approaches are considered to lead to new insights on the subject. The huge amount of data collected by instruments like the magnetic resonance tomograph increases the need for robust and reliable analytical tools. Support vector machines as discussed in this paper have become the most reknown technique to ensure feasible solutions whether for

the case of clustering or regression on the data. One very specific aspect of decision making is the question of how the individual makes a decision for which the outcome is uncertain. In these cases a valuation process takes place and a lot of research has been undertaken as for example to identify the brain regions linked to it. The focus of this thesis is to investigate the influence of applying a priori knowledge about the location of involved brain regions taken from an analysis conducted by **Mohr and Härdle et al. (2011)** to differently preprocessed data collected during the same experiment by **Mohr and Heekeren et al. (2010)**.

Motivation Support vector machines have been traditionally used to analyze data conducted by fMRI experiments [24]. And while fMRI is the forefront brain-computer interface tool [24] more sophisticated experimental designs and imaging techniques are still emerging. This situation will therefore only increase the role of statistician in the future [15]. The challenges lie in the further increasing amount of noisy data (a single fMRI acquisition in time contains information about the local brain hemodynamics at thousands of locations [5]) with a highly complicated spatio-temporal correlation structure [15]. This means that statistical tools and approaches are becoming more and more important to ensure a feasible solution to research questions in the field within a reasonable time for computation. One of the main features of statistical work in this field is finding correlation between regionally specific activations over time that may have not been previously considered and enable answering more and more complex research question as well as providing meaningful preprocessed data and robust and fast analyzing tools. The present thesis will therefore focus on one procedure of improving analysis performance through dimension reduction based on a priori knowledge about the spatial structure. This analysis will the be performed by an SVM which is in general independent of prior knowledge about hemodynamic response function [26], robust and accurate even on thousand of features [13], but whose performance can be improved by e.g. adjusting parameters or reducing the dimensionality of the feature space.

2 Methods

In the following sections I will give an introduction to the methodological procedure used to acquire and later on analyze the data.

2.1 Functional Magnetic Resonance Imaging

One of the oldest questions in neuroscience is whether the activity of a population of neurons in the brain represents aspects of an external sensor [5] and functional Magnetic Resonance Imaging is a technique developed over the last decades that seems to be the most promising approach towards answering it.

2.1.1 Introduction to fMRI

FMRI is a noninvasive method to monitor brain function with whole-brain coverage and reasonable spatial resolution [5] of approximately one millimeter. It is able to produce three dimensional images of brain activity that carry information about the blood oxygen level dependent which can serve as an indicator for brain activity in response to a stimulus. The present analysis is based on fMRI data representing changes in the subjects BOLD level over time as described later section **4.1**. "During the course of an fMRI experiment, a series of brain images are acquired while the subject performs a set of task" [15], which can vary from contemplating images to reacting to a stimulus. "To construct an image, the subject is placed into the field of a large electromagnet" [15] while a full description of the underlying physics of fMRI or MR imaging in general is not the focus of this thesis and lies beyond its scope I will give a short description here. The reader finds a brief introduction of the underlying physics in **Lindquist (2008)** and more detailed descriptions in textbooks as (e.g. **Haacke et al. (1999)**). "BOLD imaging takes advantage of inherent differences between oxygenated and deoxygenated hemoglobin. Each of these states has different magnetic properties" [15] to be measured in the experiment as the MR signal and collected for further analysis. The underlying physiological process is usually referred to as the hemodynamic response function (HRF) where the increase in inflow of oxygenated blood in the respective active area is described. **Figure 1** shows the canonical standard model of the inflow corresponding with time in seconds passed since a stimulus occurred.

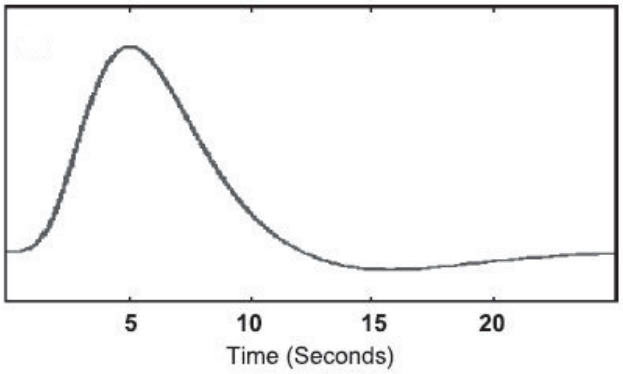

Figure 1: Standard canonical model for the HRF as used in fMRI analysis (modified from **Lindquist (2008)**)

2.1.2 General fMRI Analysis

The primary focus of fMRI data analysis seems to be finding the activation patterns in the brain representing certain mental states [13] and the general goal of a prediction algorithm is to learn a model on a given training data set in such a way that it will give a minimum error on a previously unseen data set [7]. When aligning these goals and applying such a prediction algorithm, or machine learning techniques in general, to fMRI data, the application procedure can be distinguished into four separate steps. Following **Formisano and De Martino et al. (2007)** the descriptions are visualized in **figure 2** and can be defined as follows.

Extraction The first step may be summarized as the extraction of relevant data. Here the raw data are preprocessed with respect to further analysis. The main goals are to minimize the influence of data acquisition and physiological artifacts,validate model assumptions and standardize the locations of brain regions across subjects [15]. The data may also be adapted to represent certain features depending on the goals of the further analysis.

Feature Selection Since the information about the different mental states is represented by the BOLD level as a point in a multidimensional space of voxels, the dimensionality can be very large. As it is known for machine

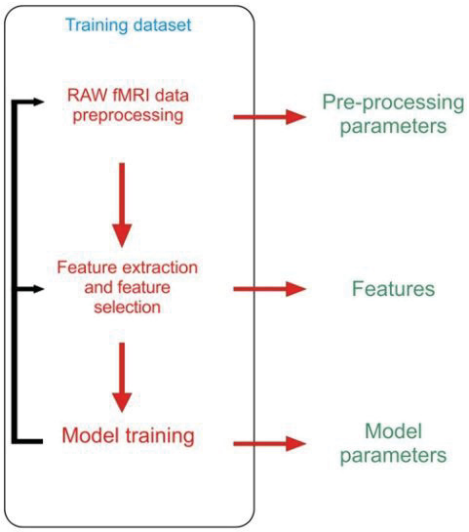

Figure 2: Main steps of a generic pattern recognition algorithm as used in fMRI data analysis (modified from **Formisano De Martino et al. (2007)**)

learning algorithms to degrade in performance when dealing with much irrelevant information, a specification of voxels may have a crucial influence on the overall performance of the analysis. A selection or reduction of the feature space is therefore crucial and will be the main focus of the present thesis. A selection to reduce feature space is therefore most advisable.

These first two steps together provide the data that can be regarded as the input for further analysis and application of machine learning algorithms. These steps are crucial to the overall performance of the later classification since these applications typically depend a great deal in the number and quality of the variables [5].

Training In the third step the actual classifier is trained on the training set, which can be just a subset of the input data. The actual method of the training and the mathematical background of a maximum *margin* classifier will be discussed in section 2.2. This step tunes the algorithm to fit the given data and requirements of the desired outcome.

Classification The trained machine is now used to classify before unseen multivariate data based on statistical regularities in the data set [5]. The

testing set is then divided into groups of participants that are significantly different between two conditions [22].

2.2 Support Vector Machines

SVMs have a long history as the most popular classifier for fMRI data analysis, both for classification and feature selection [13]. This is based especially on the SVM's ability to be trained and run on thousands of features in reasonable time [13] which distinguishes it from other classifiers. In addition SMVs are a very robust and accurate technique flexible in its application. It can be easily adapted to the problem requirements e.g. by adjusting the *kernel* function as described later. When giving an introduction on SVMs as a general technique one may consider two different cases. The first case is a SVM for a linear separable problem while the second is a non-linear SVM. I will first describe the linear separable case to give an introduction and then extend the model to fit the linear non-separable case by introducing two new features, the *slack variable* as well as the *kernel* trick. Support vector machines belong to the class of maximum *margin* classifiers. This technique is based on the idea to divide a data set into subsets by constructing a separating hyperplane in such a way, that it is furthest away from the nearest points from the opposite classes [13]. The idea is most easily described by using a graphic for the linear separable two-dimensional case.

Figure 3 shows the optimal separating hyperplane (OSH) as a line in the feature space having the maximum distance to the nearest neighbors of the different classes. The space on both sides between the closest point and the OSH is called the *margin*. The data points are correctly classified by their value for $f(x)$. To ease the formulation of following equations the values for the classification of the points will be set to $+1$ or -1.

2.2.1 Linear Separable Case

In the simplest case as shown in **Figure 3** the data points can be separated by a linear plane. Let the training data be a set $\{x_i, y_i\}$ with $i = 1, ..., l$ beeing the index of the data point from the whole training data set, $y_i \in \{-1, 1\}$ being the corresponding true membership in the positive or negative class and $x_i \in R^d$ the data point in the feature space R^d with dimension $d = 2$ here. We now assume that a separating hyperplane exists and all x_i lying on the hyperplane satisfy the condition $w^T x + b = 0$ where w is normal

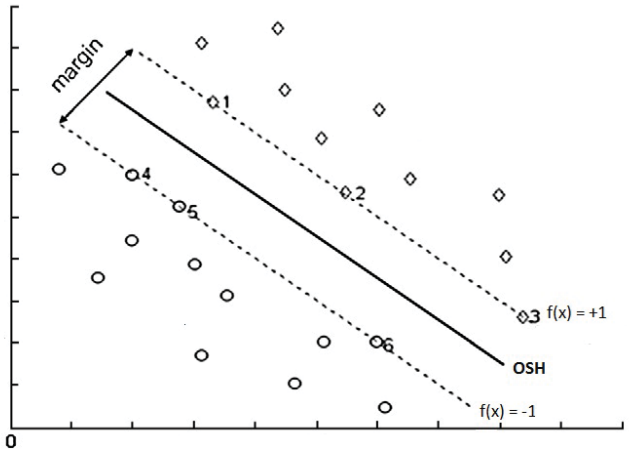

Figure 3: OSH, *margin* and support vectors for the linear separable case (modified from **Li and Liang (2009)**)

to the hyperplane, $\frac{|b|}{\|w\|}$ is the perpendicular distance from the origin to the hyperplane and $\|w\|$ is the euclidean norm of w. The width of the *margin* is defined as

$$D_+ + D_- \tag{1}$$

where $D_+(D_-)$ is the distance from the closest data point of the positive (negative) class to the hyperplane. The decision rule for the class membership of a data point X is then defined as follows: In the simple case of the linear SVM the optimization problem is reduced to finding the maximum *margin* as defined before. This leads to the constraints for all data points to be:

$$w^T x_i + b \geq +1 \tag{2}$$

for $y_i = +1$ and

$$w^T x_i + b \leq -1 \tag{3}$$

for $y_i = -1$. These constraints can be combined to form an inequality valid for all i:

$$y_i(w^T x_i + b) - 1 \geq 0, \forall i. \tag{4}$$

The decision rule towards class membership of x is then defined to be:

$$f(x) = sign(w^T x_i + b). \tag{5}$$

To define the OSH we maximize the *margin* given by the hyperplanes defined by those vectors x_i for which the inequalities from equations (2) and (3) hold true and thus lie on the hyperplanes H_1 and H_2. These are the so called *support vectors* and they are the only ones necessary to fully describe the OSH.

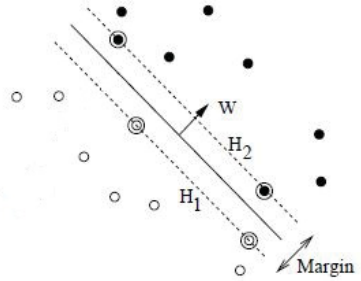

Figure 4: Finding the *margin* for linear separable data by minimizing $\|w\|^2$ (modified from **Burges (2008)**

In **figure 4** they are highlighted by a circle. For example the hyperplane $H_1 : w^T x_i + b = 1$ represents the class boundary of the positive class. **Figure 4** shows the two dimensional case for finding the OSH by maximizing the *margin*. Hence the problem of finding the maximum *margin* can be described by a minimization problem with the objective of minimizing $\|w\|^2$ with respect to the constraint from equation (4).

Lagrangian Method To allow generalization to the linear non-separable case later on and to alleviate the handling of the constraints a reformulation of the problem into a Lagrangian formulation is usually applied (see **Burges (2008), Cortez and Vapnik (1998), Christianini and Shawe-Taylor (2000)**). Since knowledge about the Lagrangian formulation of an optimization problem is assumed I will only briefly describe the basic procedure as well as the Karush-Kuhn-Tucker conditions.(for further explanation see e.g.**Fletcher (1987)**)
Introducing Lagrangian multiplier $\alpha_i, i = 1, ..., l$ for each constraint we get the following Lagrangian:

$$L_P \equiv \frac{1}{2} \|w\|^2 - \sum_{l}^{i=1} \alpha_i y_i (w^T x_i + b) + \sum_{l}^{i=1} \alpha_i. \tag{6}$$

P indicates that this is the *primal* formulation of the Lagrangian whereas D constitutes the *dual* one as described further below. Due to the constraints to be of the form $c_i \geq 0$ a positive Lagrange multiplier is applied and the constraints are subtracted from the objective function. This reduces the minimization problem to:

$$\min_{w,b} L_P \tag{7}$$

while at the same time all derivatives of the Lagrangian with respect to $\alpha_i \forall i$ will extinct whereat the constrains $\alpha_i \geq 0$ hold true for all i. The convexity of the set of constraints enables the usage of the *dual* formulation of the problem [1]. This leads to a maximization problem

$$\max L_P \tag{8}$$

with the constraints that the gradient of L_P wrt. w and b vanishes and $\alpha_i \geq 0$ holds true which leads to the following conditions:

$$w = \sum_i \alpha_i y_i x_i \tag{9}$$

and

$$\sum_i \alpha_i y_i = 0. \tag{10}$$

Substituting (9) and (10) in equation (6) leads again to a reformulation of the Lagrangian to the following:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j \tag{11}$$

Both formulation lead to the same values for w, b and α after optimization. But the *dual* formulation is much easier to compute and therefore often applied in the literature. Another note to consider is the fact that by setting $b = 0$ the constraints in equation (10) do not appear and hence reduce the number of degrees of freedom by one. The support vectors, as described above, all have a value $\alpha_i > 0$. For all other training points $\alpha_i = 0$, which means they either lie on the hyperplanes H_1 or H_2 (as shown in **figure 4**) or on their respective sides in such a way that the inequality in (4) holds true.

Karush-Kuhn-Tucker Conditions While I can not describe the procedure in total due to space restrictions and assumed knowledge of the reader I will only state the Karush-Kuht-Tucker (KKT) conditions here to show how finding a solution to the respective KKT is equivalent to solving the *SVM* problem. (see **Fletcher (1987)**)

$$\frac{\delta}{\delta w_v} L_P = w_v - \sum_i \alpha_i y_i xiv = 0, v = 1, ..., d \tag{12}$$

$$\frac{\delta}{\delta b} L_P = -\sum_i \alpha_i y_i = 0 \tag{13}$$

$$y_i(x^T y_i + b) - 1 \geq 0, i = 1, ..., l \tag{14}$$

$$\alpha_i \geq 0, \forall i \tag{15}$$

$$\alpha_i(y_i(x^T y_i + b) - 1) = 0, \forall i \tag{16}$$

While w is determined directly by the training procedure the value for b is not. But looking at the KKT conditions equation (16) reveals an easy way to compute b if one considers only the i where $\alpha_i \neq 0$. This technique is also intuitive since for $b = 0$ all hyperplanes would have to contain the origin and therefore the shift of the OSH should only be determined by the support vectors which, from the set of the training points, are the only ones affecting the OSH's shape.

2.2.2 Linear Non-separable Case

While for the linear separable case the technique stated above works fine, it will no longer deliver a feasible solution in the linear non-separable case. Therefore two new features will be introduced in the following section. First the *Slack variables* for the case of a linear SVM for linear non-separable data and second the *kernel trick* for non linear SVMs.

Slack Variables To ensure a feasible solution for the case of linear non-separable data a new cost parameter is introduced to the optimization problem. These positive *slack variables*, for the case of increasing the objective function when having a minimization problem, are given by $\xi_i, i = 1, ..., l$ and lead to an adjustment of the above stated constraint (4) as follows:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, i = 1, ..., n \tag{17}$$

while

$$\xi_i \geq 0, \forall i. \tag{18}$$

To integrate these costs for errors in the bounding feature of the hyperplanes H_i in the optimization problem one might alter the objective function to

$$\frac{1}{2} \|w\|^2 + C(\sum_i \xi_i) \tag{19}$$

where $\sum_i \xi_i$ represents an upper bound on the number of training errors and C is a weighting parameter for the influence of the penalties ξ_i. A large C therefore leads to a higher penalty for training points to be inside the margin or outside their respective boundary.

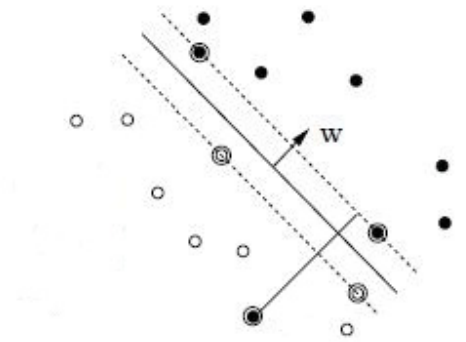

Figure 5: Finding the *margin* for a linear separating hyperplane in the linear non-separable case (modified from **Burges (1998)**)

Figure 5 shows the linear hyperplane for linear non-separable data with the distance of the hyperplane bounding the negative class on the upper right side of the OSH H_2 to the training point laying outside of the boundary $\frac{-\xi}{|w|}$. Applying these adjustments to the *dual* Lagrangian problem equation (11) stays the same but is subject to:

$$0 \leq \alpha_i \leq C \tag{20}$$

and

$$\sum_i \alpha_i y_i = 0. \tag{21}$$

This again leads to a solution for w as following:

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{22}$$

while considering again only the i for which $\alpha_i \neq 0$. For the *primal* problem the KKT conditions are used to find a solution. Due to spacial restriction I will only state the *primal* Lagrangian here but will not discuss it further. The interested reader will find more detailed information on that matter e.g. in **Burges (1998), Cortes and Vapnik (1995)**.

$$L_P = \frac{1}{2} \|w\|^2 + C \sum_i \xi_i - \sum_i \alpha_i \left\{ y_i (w^T x_i + b) - 1 + \xi_i \right\} - \sum_i \mu_i \xi_i \tag{23}$$

where μ_i are the Lagrange multipliers needed to ensure the positivity of ξ_i.

Non Linear SVM The above stated methods work only if the objective is a linear function of the data. For the other case the *kernel* trick (as introduced by **Aizerman (1964)**) is applied. To do so the training points will be mapped into a higher dimension to take advantage of the *dimensional superiority* as shown in **figure 6** for the mapping of two-dimensional data in three dimensional space. The dimension of the Euclidean space to which the data is mapped may even be infinitely large. The mapping is usually represented by a function ϕ:

$$\phi : R^{d_1} \mapsto R^{d_2} \tag{24}$$

where $d_2 \geq d_1$.

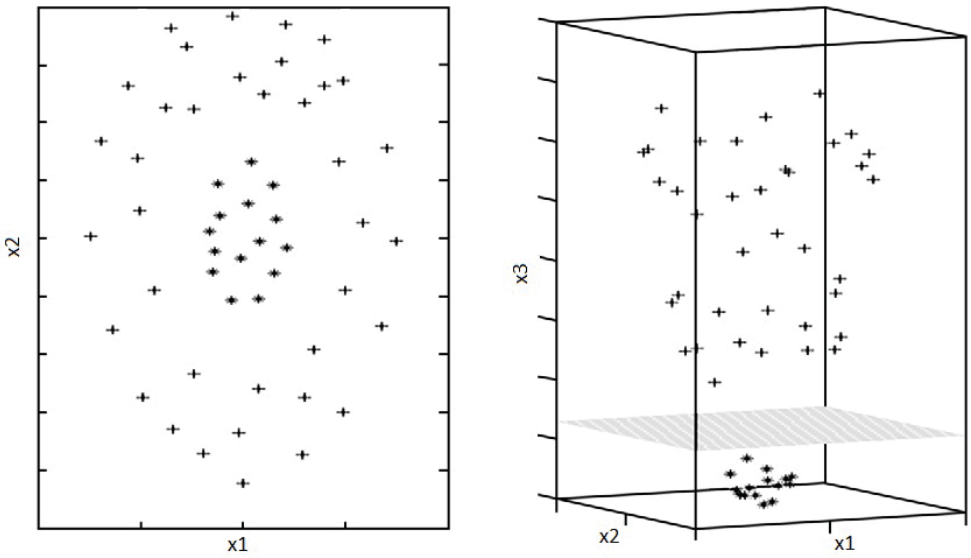

Figure 6: dimensional superiority for the case $d_1 = 2$ left side and $d_2 = 3$ on the right (modified from **Li and Liang (2009)**

In SVM related literature R^2 is often referred to as a *Hilbert* space to enable inner products not just of the form of the dot product as generally in Euclidean spaces. Since the training points appeared only in the form of dot products (e.g. see equation (10) $x_i \times x_j$) in the training process they will appear in R_2 in the same way, i.e. in form of functions as $\phi(x_i) \times \phi(x_j)$. Now the existence of a *kernel* function K is assumed to be of the form:

$$K(x_i, x_j) = \phi(x_i) \times \phi(x_j). \tag{25}$$

Therefore only K needs to be known when training the machine. This leads to a decision rule for the classification of x as the following:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)\phi(x_j) + b. \tag{26}$$

After replacing the dot product of the ϕ by the *kernel* function the decision rule becomes:

$$f(x) = \sum_{i=1}^{l} \alpha_i y_i K(x_i, x_j) + b. \tag{27}$$

While in both formulas only those i are considered again for which $\alpha_i \neq 0$, representing the support vectors. A lot of research has been conducted to find suitable *kernel* functions in the way to reduce computation time of the SVM but due to space restriction I will not go into details here. The interested reader again finds further information on that topic in (e.g. **Burges (1996)** and **Burges (1998)**). An additional note to the choice of the *kernel* is the fact that there does not exist a pair $\left\{\phi, R^{d_2}\right\}$ for every choice in such a way that a solution for the optimization problem can be found. I will only state the so called *Mercer* condition here without again going into details as described in (e.g. **Cortes and Vapnik (1995)**). A pair $\left\{\phi, R^{d_2}\right\}$ exists if and only if:

$$\int g(x)^2 dx < \infty \tag{28}$$

leading to

$$\int K(x,y)g(x)g(y)dxdy \geq 0. \tag{29}$$

This does not give a rule on how to construct ϕ but gives an idea of whether or not a *kernel* actually is an inner product in the respective space.

Gaussian radial basis function One of the most used *kernel* functions is the radial basis function(rbf) that looks like the following:

$$K(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \tag{30}$$

where the support vector is the centre of the rbf and σ determines its area of influence. A large value for σ therefore delivers a smoother decision surface and a more regular decision boundary [2]. Using a rbf is equivalent to mapping the training set into an infinite *Hilbert* space.

Another important feature of pattern recognition algorithms is the so called Vapnik Chervonenkis (VC) dimension which is a measure of the notion of capacity of a machine to learn any training set without errors [1]. It represents thereby the maximum number of points that can be shatter by a machine [2]. Since for SVMs this capacity can be very large or even infinite the VC dimension has been the focus of research but no theory has been found to guarantee high accuracy on a given problem by a family of SVMs [1].

3 The Experiment

After the introduction to the methods in general as given above I will now proceed to describe the experiment conducted by **(Mohr et al. 2010)** that delivered the fMRI data on which the later classification towards risk aversion is done. Since various studies (see **Kable Glimcher (2007)**,**Plassmann et al. (2007)**)have suggested that the value of a choice, the crucial metric in value based decision making, is represented in a network of brain regions, the present thesis tries to enhance performance of the SVM classifier by restricting analysis to a priori defined regions as described later in section 4.2. This value of a choice is again highly influenced by the decision maker's risk attitude [20]. The experimental design therefore aimed at finding neural representations of risk attitude or neural mechanisms reflecting their effect on the valuation process. In the experiment 22 subjects (age 18-35 years, 11 females) participated. All of them were native German speakers and right handed. In total five subjects had to be excluded from further analysis due to extensive head movement or modeling problems.

3.1 Task Description

The subjects were given an investment decision with uncertainty. Each trial of the Risk Perception and Investment Decision (RPID) was split in two phases. In the first phase the subjects were given a stream of ten returns of an investment. The second phase consisted of one of three task: 1.) judge the perceived risk, 2.) judge the subjective expected return or 3.) make a decision between an investment with fixed return of five percent and an investment with a return which was represented by the return stream. Each return in the stream was presented for two seconds and were drawn randomly from Gaussian distributions, an information given to the subjects in advance, with varying means (6%, 9%, 12%) and standard deviations (1%, 5%, 9%). In between these two phases was a time intervall lasting 2.5 seconds.
Figure 7 shows the sequence of a single trial of the RPID task. Each task 1.), 2.) and 3.) was performed 27 times which leads to a total of 81 trials for each subject and 1377 in total. In task 1.) subjects were asked to judge the perceived risk on a scale ranging from -5% to +15%. Task 2.) was to judge the subjective expected return on a scale from 0 (meaning no risk) to 100 (meaning maximum risk).

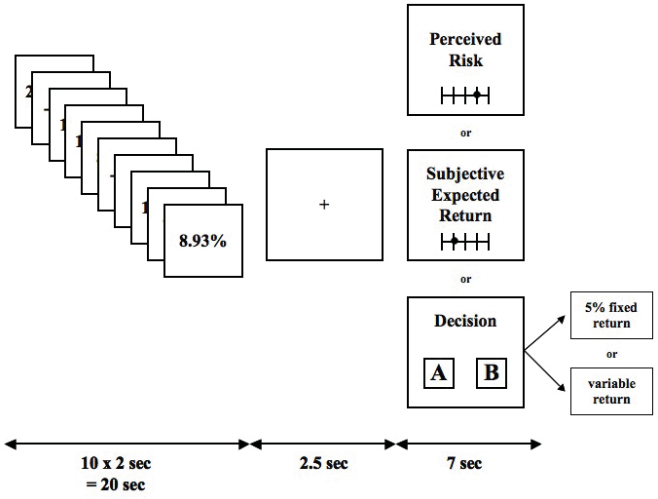

Figure 7: RPID task given to the subjects while collecting fMRI data (see **Mohr and Härdle et al. (2011)**)

3.2 Data Acquisition

The fMRI volumes were collected using a 1.5 T Magneton Sonata fMRI system (Siemens, Erlangen, Germany) with a standard head coil. Functional images were acquired using a BOLD-sensitive T2*-weighted echo-planar imaging (EPI) sequence [TR,2500 ms; echo time (TE), 40 ms; flip angle, 90; field of view, 256 mm;matrix, 64 $\times$ 64 mm; 26 axial slices approximately parallel to the bicommisural plane; slice thickness, 4mm] [20].

4 Analysis

Following the general procedure of fMRI analysis as presented in section 2.1.2 I will describe in the following the procedure of preprocessing the input data, the feature selection as well as the classification of the given data towards risk aversion.

4.1 Extraction

The first step in preprocessing the data after general motion correction etc. is the deletion of values which lie outside of the brain but are still included in

the experiments results. These values are represented by zeros in the data set. While during the conducted experiment as described in section 3 the subjects were asked to perform one of three tasks (see section 3.1) I will only consider the subjects reactions to the third question. This task is mostly correlated with the subjects risk aversion if one assumes the underlying psychological risk-return model (as described by **Mohr and Heekeren et al. 2010b**) of a decision to be represented by the following equation for the individual value $V(x)$ for an investment x..

$$v(x) = \mu(x) - \phi\sigma(x) \tag{31}$$

where $\mu(x)$ is the subjective expected return, $\sigma(x)$ the perceived risk and ϕ is the individual risk weight.
Since the canonical model of HRF is assumed (as shown in **Figure 1**) the next three values following the stimulus are most interesting to analyze. In the canonical model it is stated, that the brain reaction to a stimulus is represented in the change of the BOLD level within the next 8 seconds after the stimulus. Since the magnetic resonance tomograph used produces fMRI volumes every 2.5 seconds the most interesting points in time for the present analysis are the three following the stimulus, representing the following 7.5 seconds. To get rid of the time dimension in the data set information about these values have been merged by computing the average with respect to the initial value at the time of the stimuli. This value $\overline{fMRI}$ for the HRF was computed as described in equation (32).

$$\overline{fMRI} := \frac{1}{3}(fMRI_{t_3} + fMRI_{t_2} + fMRI_{t_1} - fMRI_{t_0}) \tag{32}$$

Where t_i represents the i-th fMRI volume collected after stimulus and t_0 the time of stimulus.

4.1.1 Variability

Following the idea, that "in addition to changes in the average BOLD signal, also the variability around this signal could carry interesting information" [22] the standard deviation is taken for these values $\overline{fMRI}$.

$$\Delta(\overline{fMRI}) = std(\overline{fMRI}) \tag{33}$$

These values $\Delta(\overline{fMRI})$ are then used as representatives for each voxel's reaction to the stimulus. A strong simplification of the underlying HRF

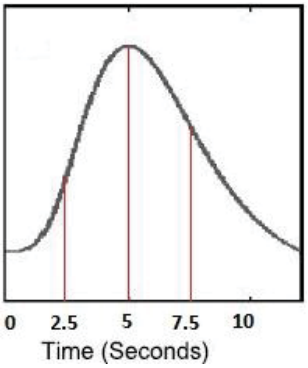

Figure 8: The interesting points in time after stimulus as values of the hemodynamic response function

is visualized in **figure 8**. The resulting data, that will later on be used for classification towards the subjects risk attitude, therefore represent the information hold in the variability of the respective BOLD signals for each subject for the third question.

4.1.2 ROI Visualization

To ensure only those areas of the brain are chosen in the feature selection step, that correspond with the given task of the experiment as described above, the values for the respective indices of the data vector representing $\Delta\overline{fMRI}$ had to be known a priori. These values were collected by the help of the *nii* package for $MATLAB^{®}$ software.The graphics shown in **figures 9 to 13** were produced by applying quantile selection to the factor loadings of the dynamic semi-parametric factor modeling (DSFM) (see **Mohr and Härdle et al. (2011)**). The 0.0005 quantile and the 0.9995 quantile were chosen and the values of the factor loadings were increased significantly to amplify the contrast in the *nii* image thereby making visual distinction between ROI and non related brain areas possible. The values to be visualized in the nii images were taken from the results of the dynamic semi-parametric factor modeling as conducted by **Mohr and Härdle et al. (2011)**. A specification of the factor loadings to represent the several areas of interest to be feasible as participating in the investment decision were given through the results presented in **Mohr and Härdle et al. (2011)**.

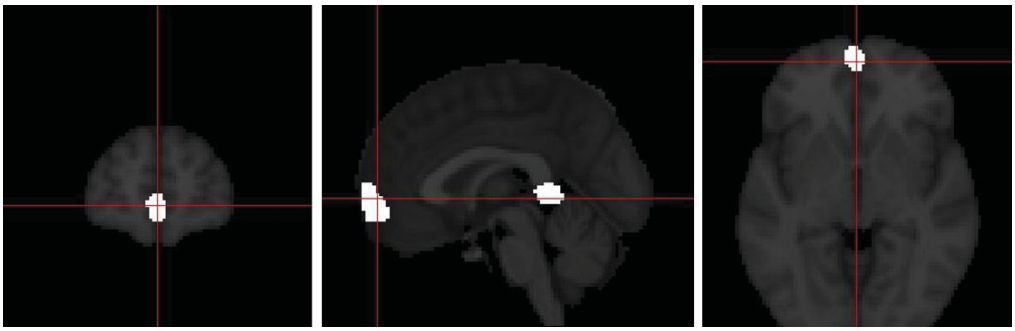

Figure 9: ROI VMPFC represented by factor loading $\widehat{M}_2$

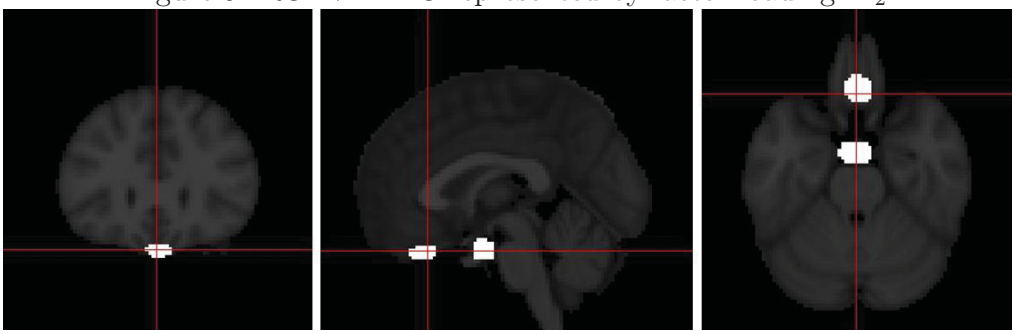

Figure 10: ROI MOFC represented by factor loading $\widehat{M}_4$

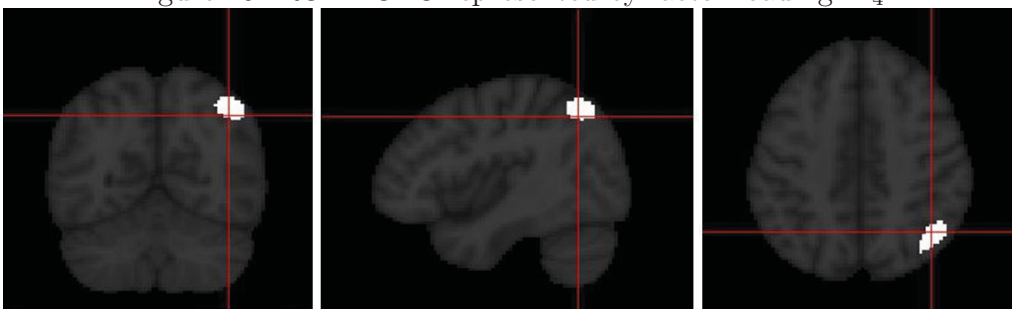

Figure 11: ROI LPC represented by factor loading $\widehat{M}_{13}$

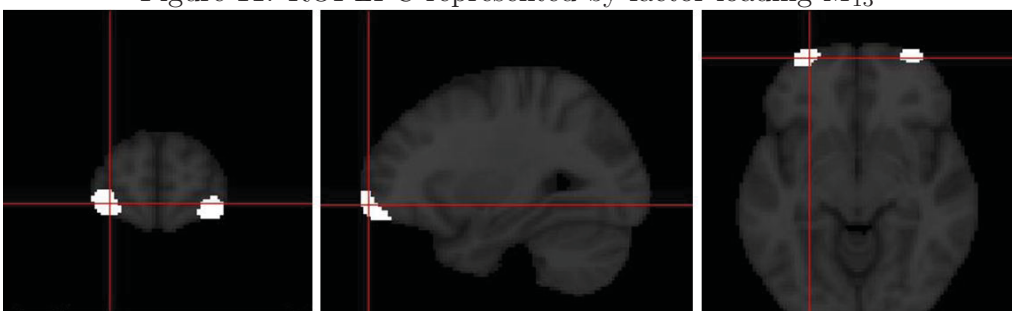

Figure 12: ROI LOFC represented by factor loading $\widehat{M}_{14}$

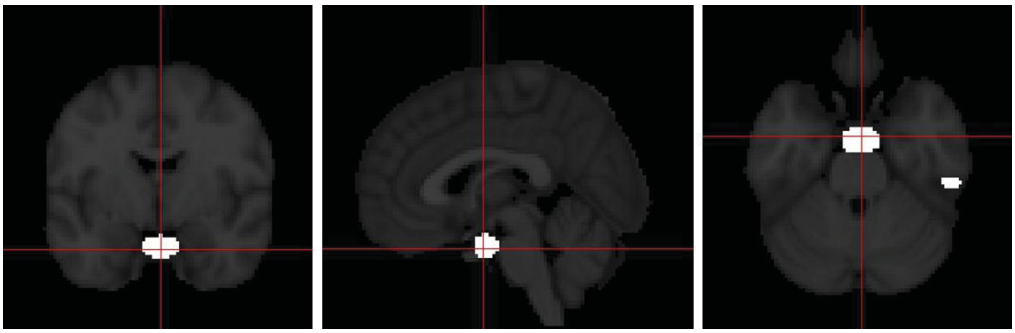

Figure 13: ROI AMG represented by factor loading $\widehat{M_3}$

4.2 Feature Selection

Using a preset number of voxels, as proposed for example in [24], can significantly increase overall performance of classification and is therefore performed here. Generally in fMRI data analysis for classification feature selection is equivalent to voxel selection [5]. To further reduce computation time the given data matrix for the $\Delta\overline{fMRI}$ was reshaped as a vector representing only the areas of the brain considered relevant. This vector had to consist only of the data of the matrix as defined by a priori known indices. These indices were taken by accounting for the voxel indices of the highlited areas in the above created *nii* images. Two requirements were formulated for index selection. First, to ensure differences in location and shape of the ROIs between subjects do not have a strong impact on the classification, the chosen set of voxels had to be minimal, to lie within the respective brain area of all subjects. Second the respective brain area highlighted by the factor loadings had to be known to be associated with decisions under risk. Following suggestions in the literature and the results of [20] the respective areas are: *ventromedial prefrontal cortex (VMPFC), medial orbitofrontal cortex (MOFC), lateral orbitofrontal cortex (LOFC), lateral parietal cortex (LPC), amygdala (AMG).*
Table 1 gives an overview of the selected voxels for further analysis by showing the range of voxels for each dimension for each ROI.

This approach not only decreases computation time but can, for the case of areas carrying redundant information or have already been considered, which would add little to accuracy of classification, keep those redundancies from seriously impairing performance [5]. I therefore reduced the size

ROI	X axis	Y axis	Z axis
VMPFC	43:48	91:96	35:40
MOFC	43:46	76:82	21:26
LPC	65:70	30:36	59:64
LOFC	29:33	92:96	30:34
AMG	48:52	61:65	25:29

Table 1: Regions of interest with respective index range in units of voxel

of the input data for the classifier to solely represent the ROIs as described above. The areas selected, as proposed in recent studies (e.g.: **Mohr and Heekeren et al. (2010a), Mohr and Härdle et al.(2011), Mohr and Heekeren et al. (2010c)**) are of the following sizes as described in **table 2**. The reshaping of the matrix was done applying the MATLAB® reshape

ROI	Dimension
VMPFC	$6 \times 6 \times 6$
MOFC	$4 \times 7 \times 6$
LPC	$6 \times 7 \times 6$
LOFC	$5 \times 5 \times 5$
AMG	$5 \times 5 \times 5$

Table 2: Regions of interest with respective dimension in units of voxel

function to the given variable representing the standard deviation of the subjects $\Delta(\overline{fMRI})$. For each ROI the matrix containing the information about $\Delta(\overline{fMRI})$ was reduced to only contain the information at the given index of the respective area. These matrices representing the values for each ROI were then joined to create a vector containing all the information of the subjects $\Delta(\overline{fMRI})$ for all ROIs. By the given dimensions as stated in table 2 this led to a 886 dimensional vector that was used as input for the following classification. The basic idea of the voxel selection was to find indices representing enough voxels of the ROI to contain sufficient information for the classification but at the same time representing a small enough space to be applicable to all subjects regardless of inter subject differences concerning the exact area of the ROI due to minor head movement or other causes. The reshaping therefore created a much smaller vector representing adequate information about $\Delta(\overline{fMRI})$ to be used for classification. This approach fol-

lows the intention that a priori reasons for a given set of criteria for voxel selection will provide information about a given discrimination [5].

4.3 Training

The training of the classifier was done using the so called double cross-validation method. This approach reduces the risk of overparameterization and too optimistic results [20]. It consists of 4 steps that can be described by the following pseudo code.

1. `(leave-one-out procedure) to separate data set into training set and testing set`
2. `leave one out cross-validation on training set to compute decision rule for the SVM`
3. `use decision rule from 2. to classify the training set.`
4. `repeat 1-3 for all possible distinct training & testing set separations`

This procedure was applied on the data matrix containing the vectors from each subject that represent the $\Delta(\overline{fMRI})$ for the respective ROI as described above (section 4.2). From this matrix for each iteration of the algorithm one subject's information was taken aside to represent the testing set. The other rows of the data matrix, together with the true class labels, were then used to define the parameters for the decision rule of the SVM. This was achieved by again using the leave-one-out procedure to separate the data in two groups. The first group consists of one vector representing only one subject's information and the second group of the vectors for all other subjects. This way the parameter could be optimized for the decision rule as described in section 2.2. A more detailed description of the procedure exceeds the scope of this thesis and can be found in **Mohr and Härdle et al. (2011)**.

4.4 Implementation

The training as well as the previous feature selection step were executed in MATLAB®. The vectors representing the subjects $\Delta(\overline{fMRI})$ were of the size

[886, 1]. It was further necessary to join these vectors into one data matrix K of the dimensions $[17, 886]$ to serve as input for the support vector machine as implemented in the MATLAB® Bioinformatics Toolbox™. This function *svmtrain* was further specified by the definition of several parameters such as the respective *kernel* function. The chosen *kernel* function was a Gaussian radial basis function as described exemplarilly under 2.2.2 by equation (30).

4.5 Results

The presented results were obtained by applying the algorithm as described above. In addition parameters were changed during training iterations. The *kernel* parameters that were changed are the values for σ and C. These values lied in the intervals as defined in **table 3**.

	C	σ
min	0.01	0.01
max	80	1.2
divided by	1.6	0.8
range	50	15

Table 3: Ranges for the values capacity C and σ of the rbf

The classification rate (CR) resulting from this algorithm is given by the ratio of truly classified subjects over the number of all subjects N.

$$CR = \frac{\sum TRUE_j}{N} \tag{34}$$

where j is the class index for weak and strong risk avers subjects.

N=17	Strong	Weak
real	10	7
estimation	17	0
TRUE	10	0
CCR	100%	0%

Table 4: Correct classification rate (CCR) for strongly and weakly risk avers subjects

Table 4 displays the real and estimated values for the class sizes and the

number of truly classified subjects per class ($TRUE$). The CR for this approach is therefore equal to $\frac{10}{17} = 0.5882$ or 58.82%. This is equal to $1 - rate$ where $rate$ is defined as the mean of the average errors per training iteration:

$$rate(C, \sigma) = mean(\overline{er}_i) \tag{35}$$

for all subjects i.
These average errors $\overline{er}_i$ for classifying are displayed in **table 5**.

1	2	3	4	5	6	7	8
0.4375	0.3750	0.4375	0.4375	0.3750	0.3750	0.4375	0.3750
9	10	11	12	13	14	15	16
0.4375	0.3750	0.3750	0.4375	0.4375	0.4375	0.4375	0.4375
17							
0.3750							

Table 5: Average error ($\overline{er}_i$) for subjects 1 to 17

The $rate(C, \sigma)$ is equal to 0.4118 for all values of C and σ. This resulted again in an overall CR of 58,82%. These results did not change during the iterations of the algorithm. Which means that for all values of C and σ all subjects were classified as strongly risk averse. Since these results are not in a satisfactory manner I rerun the algorithm with an altered input vector for the classifier. The new input represents the mean of the previously defined ROIs.

$$K = mean(ROI(\Delta\overline{fMRI})) \tag{36}$$

where K is the input vector and ROI is the vector containing the previously defined brain areas representing the respective values for $\Delta\overline{fMRI}$. This review of the algorithm resulted in different values for CR, $\overline{er}_i$, the $rate$ and was able to distinguish between strongly and weakly risk avers subjects. The best CR reached by this approach is as well 58,82%. But the rate for correctly classifying the subjects (CCR) to the classes varied based on the given values for C and σ. **Table 6** shows examples for the best classification rates towards strongly and weakly risk averse classes for the given values for C and σ.
These results display the dependence of the performance on the previously defined *kernel* parameters. To get information about the average performance of the technique I computed the average CR by taking the mean of

N=17	Strong	Weak
C	all values	0.01
σ	0.01	1.13
CCR	100%	58.04%

Table 6: Best CCR for strongly and weakly risk avers subjects based on mean of ROI

rate and setting it to be:

$$\overline{CR} = (1 - mean(rate(C, \sigma))) \times 100\%. \tag{37}$$

This method thereby achieved an average classification rate $\overline{CR}$ of 48.35%.

5 Interpretation and Comparison

The given results in section 4.5 show that the algorithm was not able to distinguish between strongly and weakly risk avers subjects when using just the defined ROI as input. But a reduction of the dimensionality of the input vector by taking the mean of the respective ROI enabled the algorithm to recognize weakly risk averse subjects. While the overall performance measured by the classification rate CR did not improve the fact that a separation took place is evidence for the main cause of the weak performance of the SVM if just using the high dimensional ROI as input.

5.1 Interpretation of Results

The presented results therefore underline the importance of a well conducted feature selection step as described above. The reduction of the dimensionality of the input vector illustrated the dependence of the SVM on the quality of the input data. Since the main task of the present study was to investigate the influence of the feature selection step via a priori defined regions of interest on the overall performance the given results represent good evidence towards the statement that reducing the input data to just the respective ROI in connection with the voxel's representation of the $\Delta\overline{fMRI}$ can provide a classification towards the subjects risk aversion. On the other hand this study showed that the quality of the input data, especially its dimensionality, play a crucial role in the overall performance of the trained machine.

5.2 Comparison

A comparison with other fMRI studies is difficult but due to the shared underlying experiment with the paper by **Mohr and Härdle et al. (2011)** seems at least plausible for this case. While the CR achieved in the present study lies under the one achieved in that paper, the computation time is significantly smaller. This might of course also be tracked back to the fact that the data size is smaller and I could already use the results of the DSFM approach to define respective voxel location for the definition of the ROIs.

6 Outlook

The given classification may be further improved by optimizing the values for the *kernel* parameters C and σ as well as optimizing the input data. The general methodology of reducing the input by predefining ROIs proved to deliver solutions in reasonable computation time and should therefore be considered for further research. One possible future approach is to redefine the procedure of determining the voxel's locations for the respective ROI. This should include information about the shape of the ROI considering the high correlation between neighboring voxels and the idea that information is not just entailed in the maximally responsive regions [7]. This represents a more promising approach for answering the question about the influence of the activity of a population of neurons in the brain, since these voxel-by-voxel approaches only look at the "tip of the iceberg" [7] of the information included in the response patterns.

References

[1] Burges, C., A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge discovery*, Vol. 2, p. 121-167, 1998

[2] Chin, K., *Support Vector Machines applied to Speech Pattern Classification*, Dissertation submitted to University of Cambridge, Sept. 1998

[3] Christianini, N., Shawe-Taylor J., *Support Vector Machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, UnitedKingdom, 2000

[4] Cortes, C., Vapnik, *The nature of statistical learning theory.* New York [et al.] Wiley, 1998

[5] Cox, D., Savoy, R., Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex, *NeuroImage*, Vol. 19, p. 261-270, 2003

[6] Fletcher, R., *Practical Methods of Optimization*, John Wiley and Sons, Inc., 2nd edition, 1987.

[7] Formisano, E., De Martino, F., Valente, G., Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning, *Magnetic Resonance Imaging*, vol. 26, p. 921-934, 2008

[8] Hafner, C., Haerdle, W., Prastyo, D., Support Vector Machines with Evolutionary Feature Selection for Default Prediction,*SFB 649 Discussion Paper 2012-030*, 2012

[9] Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning; Data Mining, Inference and Prediction* Springer Series in Statistics,Springer Science+Business Media, 2nd edition, 2009

[10] Homepage of MathWorks Documentation Center for SVMStruct function in MATLAB®, *http://www.mathworks.de/de/help/bioinfo/ref/svmtrain.html*, last called November 26, 2012

[11] Homepage of nii package for MATLAB®, *http://www.rotman-baycrest.on.ca/ jimmy/NIfTI/*, last called November 26, 2012

[12] Hsu, C., Chang, C., Lin, C., A Practical Guide to Support Vector Classification, *Technical Report, Department of Computer Science, National Taiwan University*, 2003

[13] Kuncheva, L., Rodrigeuz,J., Classifier ensembles for fMRI data analysis: an experiment, *Magnetic Resonance Imaging*, vol 28, p. 583-593, 2010

[14] Lia, H., Liang, Y., Xub, Q., Support vector machines and its applications in chemistry, *Chemometrics and Intelligent Laboratory Systems*, Vol. 95, Issue 2, p. 188-198, 2009

[15] Lindquist, M., The Statistical Analysis of fMRI Data, *Statistical Science*, Vol. 23, No. 4, p. 439-464, 2008

[16] Luts, J., Ojeda, F., Van de Plas, R., et al., A tutorial on support vector machine-based methods for classification problems in chemometrics, *Analytica Chimica Acta*, Vol. 665, Issue 2, p. 129-145, 2010

[17] Mitchell, T., Computational Models of Neural Representations in the Human Brain, (extended abstract) *DS 2008, Lecture Notes in Artificial Intelligence 5255,* J.-F. Boulicaut, M.R. Berthold, and T. Horvarth (Eds.), Springer-Verlag Berlin Heidelberg, p. 2627, 2008.

[18] Mohr, P., Biele, G., Heekeren, H., Neural Processing of Risk, *Journal of Neuroscience*, Vol. 30(19), p. 6613-6619, 2010a

[19] Mohr, P., Biele, G., Krugel, L., Li, S., Heekeren, H., Neural foundations of risk-return trade-of in investment decisions. *NeuroImage*, Vol. 49, p. 2556-2563, 2010b.

[20] Mohr, P.,Härdle, W.,Mysickova, A., Song Song, Majer, P., Risk Patterns and Correlated Brain Activities.Multidimensional statistical analysis of fMRI data with application to risk patterns, submitted to *Psychometrika*, 2012

[21] Mohr, P., Li, SC.,Heekeren, HR., Neuroeconomics and aging: neuromodulation of economic decision making in old age.*Neuroscience and Behavioural Reviews*, vol. 34, p. 678-688, 2010c

[22] Mohr, P., Nagel, I., Variability in brain activity as an individual difference measure in neuroscience? *The Journal of Neuroscience* 30, 2010

[23] Murphy, K., *MACHINE LEARNING: A PROBABILISTIC APPROACH* not yet printed, Draft of August 11, 2010

[24] Pereira, F., Mitchell, T., Botvinick, M., Machine learning classifiers and fMRI: a tutorial overview., *NeuroImage*, vol 45, p. 199-209, 2009

[25] Wallisch, P., Lusignan, M., Benayoun, M., Baker, T., Dickey, A., Hatsopoulos, N., *MATLAB® for Neuroscientists, An Introduction to Scientific Comupting in MATLAB®*, Academic Press, 2009

[26] Wang, Ze, A hybrid SVM-GLM approach for fMRI data analysis, *NeuroImage*, vol. 46, p. 608-615, 2009

Declaration of Authorship

I hereby confirm that I have authored this Bachelor thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Berlin, 26.11.2012 *Patrik Bey*